Model-assisted Estimation of Mixed-Effect Model Parameters in Complex Surveys

Eric V. Slud, U.S. Census Bureau, CSRM
& Univ. of Maryland, Math. Dept.

Joint Statistical Meetings, July 2019

Disclaimer

This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are the author's and not necessarily the Census Bureau's.

Outline

• Problem definition, specialized to 2-level models in complex surveys

• Previous research – assumptions & theoretical results

• New Pseudo-likelihood EM method – exposition and simulation results in 2-level linear ANOVA model

• Generality of available methods – Further models & examples

Random Effects Models in Complex Surveys

Problem Formulation

- existence of design- and model-consistent estimator of multilevel-model parameters in complex surveys with many independent (ultimate) clusters including random effects

shared cluster effects make survey-weighted (pseudo) loglikelihoods not directly applicable

- existence of consistent method-of-moments estimators

- existence of other (estimating-equation-based) consistent method-of-moments estimators

- **Key issue –validity of estimation methods for both** *non-informative* **and** *informative* **weights**

Multilevel Survey Superpopulation Framework

Survey frame $\mathcal{U}$, records $\{y_i, \mathbf{z}_i\}_{i \in \mathcal{U}}$, probability sample $\mathcal{S} \subset \mathcal{U}$ with inverse single-inclusion (conditional) prob. weights w_i

Multilevel: population units are multiply (here doubly-) indexed $i = (j,k)$ where $k(i)$ denotes cluster, $\mathcal{U}_k = \{i = (j,k) : \; k = k(i)\}$
Assume sample hierarchical with cluster sampling weights ω_k, within-cluster weights $w_{j|k} \equiv w_{(j,k)}/\omega_k$

Superpopulation model $\{(y_i, \mathbf{z}_i) : \; i \in \mathcal{U}_k\}$ independent & satisfy

$$y_{(j,k)} \overset{indep}{\sim} f(y \,|\, z_{(j,k)}, a_k, \beta, \eta_1), \qquad a_k \overset{indep}{\sim} g(a, \eta_2), \qquad \theta = (\beta, \eta_1, \eta_2)$$

Noninformative sampling (of clusters/units) if $\{(y_i, \mathbf{z}_i) : \; i = (j,k) \in \mathcal{S} \cap \mathcal{U}_k\}$ satisfies same model, for $k \in \mathcal{S}_C = \{k(i) : \; i \in \mathcal{S}\}$

Background — Previous Work, Quick Summary

With noninformative sampling: consistent estimation can ignore survey weights. What about informative sampling of clusters ?

Binder (1983) & Skinner (1989) showed that pseudo-likelihood $\sum_{i \in S} w_i \log f(Y_i \,|\, \mathbf{Z}_i, \theta)$ provides valid inference under independent-unit parametric superpopulation model even under informative (outcome-data-biased) sampling

Pfeffermann et al. (1998) considered informatively sampled linear (2-level ANOVA) model

$$y_{(j,k)} = \beta' z_{(j,k)} + a_k + \epsilon_{(j,k)}, \;\; a_k \sim \mathcal{N}(0, \sigma_a^2), \;\; \epsilon_{(j,k)} \sim \mathcal{N}(0, \sigma_e^2)$$

with complicated iterative WLS procedure involving weight-rescaling. No proofs given; method apparently works with noninformative sampling (in their and Korn & Graubard's simulations).

Background Summary, Linear Models Cont'd

Korn & Graubard (2003) showed in case with no covariates $z_{(j,k)}$ ($\beta = \mu$): Pfeffermann et al. methods not consistent for general informative sampling; K & G provided consistent method-of-moments method based on joint inclusion probabilities.

Asparouhov (2006) amplified weight-scaling idea, showing consistency in some informative-sample cases; appealed to same 'pseudo-logLik' as Rabe-Hesketh & Skrondal (2006), below.

Special Role of Linearity

With informatively sampled clusters, linearity enables consistent estimation via WLS and residual moments:

$$\widehat{\beta}_{\text{WLS}} = \Big(\sum_{(j,k)\in\mathcal{S}} w_{(j,k)} \mathbf{z}_{(j,k)}^{\otimes 2} \Big)^{-1} \sum_{(j,k)\in\mathcal{S}} w_{(j,k)} \mathbf{z}_{(j,k)}\, y_{(j,k)}$$

$$\widehat{\sigma}^2_{e,\text{Mom}} = (\sum_{k\in\mathcal{S}_C} \omega_k)^{-1} \sum_{(j,k)\in\mathcal{S}} \omega_k \, \text{var}(\widehat{e}_{(j,k)} : (j,k) \in \mathcal{S})$$

$$\widehat{\sigma}^2_{a,\text{Mom}} = (\sum_{(j,k)\in\mathcal{S}} w_{(j,k)})^{-1} \sum_{(j,k)\in\mathcal{S}} w_{(j,k)} \, \widehat{e}^2_{(j,k)} \; - \; \widehat{\sigma}^2_{e,\text{Mom}}$$

$$\widehat{e}_{(j,k)} \; = \; y_{(j,k)} \; - \; \widehat{\beta}'_{\text{WLS}} \, \mathbf{z}_{(j,k)}$$

Background Summary, General Models

Rabe-Hesketh and Skrondal (2006): maximize $logLik =$

$$\sum_{k \in \mathcal{S}_C} \omega_k \log \int \exp\Big(\sum_{j \in \mathcal{S}_k} w_{j|k} \log f(y_{(jk)} \,|\, \mathbf{z}_{(jk)}, a_k, \beta, \eta_1) \Big)\, g(a_k, \eta_2) da_k$$

But integral expression is not a likelihood, and consistency of estimation is justified only when (all) cluster-sizes go to ∞.

Rao, Verret and Hidiroglou (2013) generalize Korn & Graubard's method of moments, estimating consistently based on composite pairwise likelihoods weighted by joint inclusion probabilities.

Pseudo-EM Method

Census augmented logLikelihood

$$\sum_k \log g(a_k, \eta_2) + \sum_{(j,k)\in\mathcal{U}_k} \log f(y_{(j,k)}|\mathbf{z}_{(j,k)}, a_k, \beta, \eta_1)$$

is estimated design-consistently (for augmented survey dataset and all parameters θ) by $l_w(\theta) =$

$$\sum_{k\in\mathcal{S}_C} \omega_k \log g(a_k, \eta_2) + \sum_{(j,k)\in\mathcal{U}} w_{(j,k)} \log f(y_{(j,k)}|\mathbf{z}_{(j,k)}, a_k, \beta, \eta_1)$$

As for usual EM algorithm, but now using estimated log-likelihood, iteratively for initial θ_0,

$$\theta_1 = \arg\max_\theta \; E_{\theta_0}\Big(l_w(\theta) \,|\, I_{[(j,k)\in\mathcal{S}]}, w_{(j,k)}, y_{(j,k)}, \mathbf{z}_{(j,k)} \Big)$$

Implementation & Theory for Pseudo-EM

Need to be able to compute conditional distributions for a_k in last E-step. For this, generally need noninformative sampling within clusters, with weights $w_{j|k}$ free of $y_{(j,k)}, a_k$.

When this holds, under general asymptotic conditions (also related to EM convergence and unique MLE or local starting values), convergent pseudo-EM maximizer is approximately the census-logLik MLE.

Special Case of Linear ANOVA Model

(1) When within-cluster sampling is noninformative, explicit conditional distributions $a_k \sim \mathcal{N}(\gamma_k(\bar{y}_{\cdot,k} - \beta'\bar{\mathbf{z}}_{\cdot,k}),\, (1-\gamma_k)\sigma_a^2)$ (where $\gamma_k = n_k\sigma_e^2/(n_k\sigma_e^2 + \sigma_a^2),\ n_k = |S_k|$) lead to explicit EM iterations $\theta_0 \mapsto \theta_1$ in terms of weighted survey data.

(2) When $y_{(j,k)} = \mu + a_k + \epsilon_{(j,k)}$, and weights are constant within cluster, pseudo-EM estimator is **identical** to WLS and residuals-based estimators $\widehat{\mu}_{\text{WLS}},\, \widehat{\sigma}^2_{a,\text{Mom}},\, \widehat{\sigma}^2_{e,\text{Mom}}$. Analogous result holds in regression ANOVA when $\mathbf{z}_{(j,k)}$ are constant across j.

(3) When sampling within-cluster is noninformative, pseudo-EM and WLS & residual-MOM estimators remain extremely close and consistent, as confirmed by simulations.

Linear Regression ANOVA, cont'd

(3) In some settings with informative within-cluster sampling, pseudo-EM still does remarkably well; e.g., where a noninformative sample is modified as in Korn and Graubard by subsampling with prob. 1/2 those units with $|\epsilon_{(j,k)}| > 0.6745\,\sigma_e$, based on 1000 iterations, in samples of $\approx$ 500 clusters of size $\approx$ 24 from a population of $2 \cdot 10^6$) the average parameter estimators were

	β_0	β_1	σ_a^2	σ_e^2
PseudoEM	-0.0124	1.0014	0.9946	0.9816
WLS/Mom	-0.0084	1.0039	1.2748	0.7320
True	0	1	1	1

Further Research on this Topic

In other (nonlinear) models, only pseudo-EM provides consistent estimators based on complex surveys with informatively sampled clusters in terms of single-inclusion probability weights, even if sampling within clusters is noninformative:

(i) Beta-binomial with random effects:

$y_{(j,k)} \sim \text{Binom}(\nu_{jk}, \pi_k)$, $\pi_k \sim \text{Beta}(\tau\mu,\, \tau(1-\mu)$ *iid*

(ii) Logistic regression with random effects:

$y_{(j,k)} \sim \text{Binom}(\nu_{jk}, \text{plogis}(\beta' \mathbf{z}_{(j,k)} + a_k))$, with $a_k \sim \mathcal{N}(0, \sigma_a^2)$ *iid*

(iii) Nonlinear regression: $y_{(j,k)} = h(\beta' \mathbf{z}_{(j,k)} + a_k) + \epsilon_{(j,k)}$

Extensions, continued

In these model settings (i) still allows explicit conditional distributions and EM iterations. In (ii) and (iii), the E-step must be implemented numerically, with an approach such as adaptive Gaussian Quadrature (Pinheiro & Bates 1995).

References

Binder, D. (1983), *Internat. Statist. Review*

Korn, E. and Graubard, B. (2003), *Jour. Royal Statist. Soc.* B

Pfeffermann, D., Skinner, C., et al. (1998), *JRSS* B

Rabe-Hesketh, S. and Skrondal, A. (2006), *JRSS* A

Rao, JNK et al. (2013), *Survey Methology*

Thank you !

Eric.V.Slud@census.gov